Herramientas Econometricas
CEF y Proyección Lineal

Pasquini, Ricardo

IAE Universidad Austral

September 9, 2024

CEF

- Motivación: Descripción del ingreso en la población
- Funcion de Esperanza Condicional (CEF) y sus propiedades
- Varianza Condicional y Varianza del Error

Ejemplo

Analizaremos la teoría junto al caso de los ingresos individuales en CABA.

Distribuciones Poblacionales

- Supondremos Y es una *variable aleatoria* proveniente de una población con una función de densidad acumulativa (CDF)

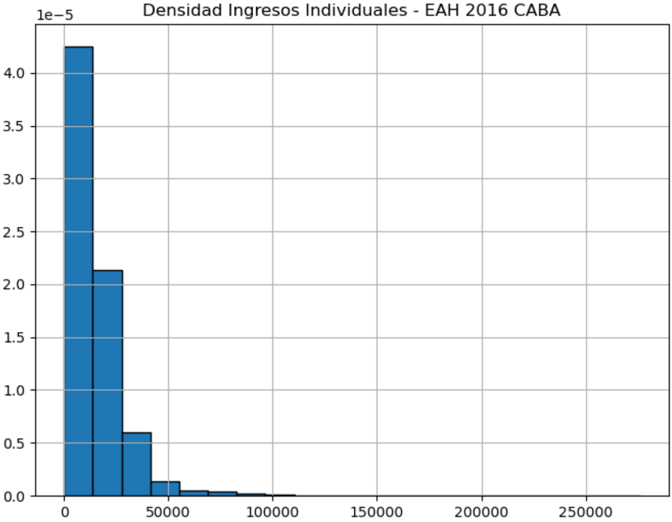
$$F(y) = Prob(Y \leq y)$$

- Supondremos factores explicativos como $X_1, X_2, \ldots, X_k$ tambien como variables aleatorias con sus respectivas distribuciones.

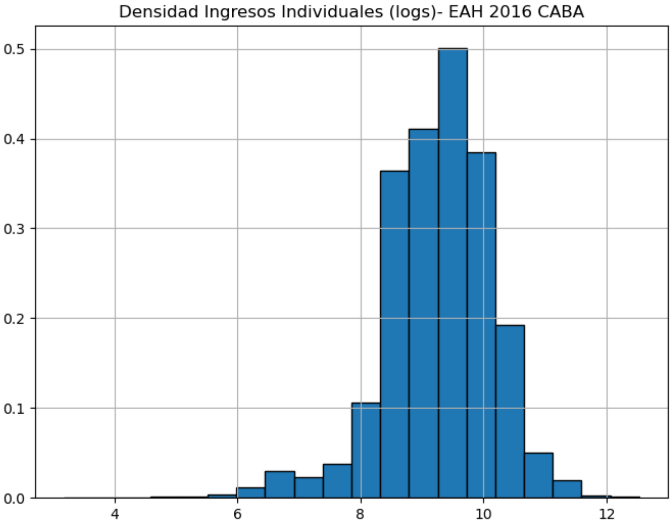
Repaso Estadístico

- Una *variable aleatoria* es una variable cuyo resultado es a priori desconocido y queda determinada por un experimento
- En este caso las variables provienen de la distribución de la población.

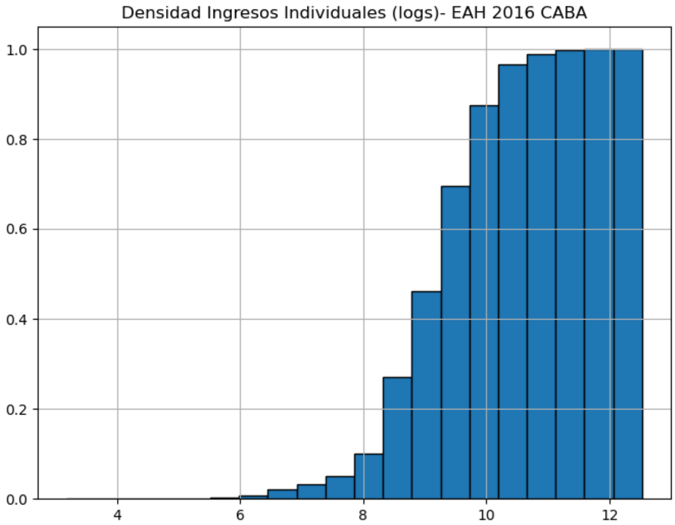
Distribución del ingreso - Densidad



Distribución del ingreso - Densidad



Distribución del ingreso- Densidad Acumulada

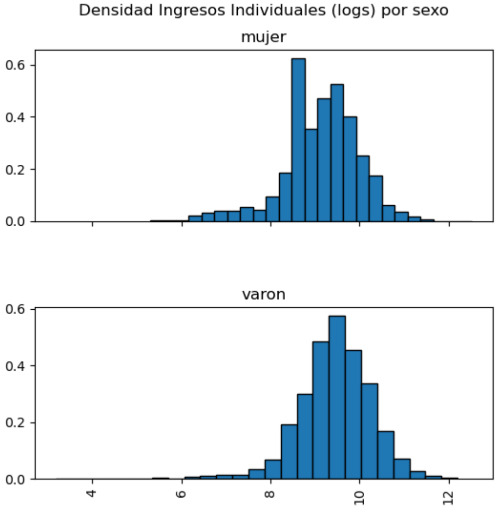


Distribucion del ingreso por sexo

¿Explica el sexo la distribución del ingreso? ¿Varía la distribución del ingreso de acuerdo al sexo?

- Lo inspeccionaremos gráficamente
- Utilizaremos el valor esperado condicional como una aproximación

Distribucion del ingreso por sexo



Distribucion del ingreso por sexo

Aproximación: Promedio muestral

- Intuitivamente un valor que sirve para describir la muestra es el promedio (Avg)

$$Avg[Y|sexo = "hombre"] = 9.44$$

$$Avg[Y|sexo = "mujer"] = 9.13$$

	count	mean	std	min	25%	50%	75%	max
sexo								
mujer	5324.000000	9.136385	0.877134	5.298317	8.612503	9.210340	9.718660	12.528156
varon	4789.000000	9.444136	0.813080	3.178054	8.987197	9.472705	9.928180	12.206073

Table: Descripcion de Ingresos

Distribucion del ingreso por sexo

Aproximación: Esperanza Condicional

- Nuestro objetivo no es solo describir una muestra. Queremos establecer una teoría a nivel poblacional.
- Proponemos la esperanza matemática

$$E[Y|sexo = "hombre"]$$

$$E[Y|sexo = "mujer"]$$

Repaso Estadístico

- Si X es una variable aleatoria, el valor esperado (o expectativa) de X, denotado $E[X]$ y a veces μ_X o simplemente μ, es un promedio ponderado de todos los valores posibles de X. Los pesos están determinados por la función de densidad de probabilidad.
- Suponiendo un caso discreto donde X toma k posibles valores

$$E[X] = \sum_{j=1}^{k} X_k P(X_k)$$

Repaso Estadístico

- Suponiendo un caso continuo tenemos

$$E[X] = \int_{-\infty}^{\infty} Xf(X)dx$$

1. Para cualquier constante

$$E[c] = c$$

2. Para cualquier constante a y b

$$E[aX + b] = aE[X] + b$$

3. Si $\{a_1, a_2..a_k\}$ y $\{X_1, X_2..X_k\}$

$$E[\sum a_i X_i] = \sum a_i E[X_i]$$

Función de Esperanza Condicional (CEF)

- En general, es natural que estemos interesados en el ingreso *de las mujeres* , de los hombres, es decir para un subgrupo.
- En terminos estadísticos es natural que estemos interesados en conocer el valor esperado de una variable Y condicional a un cierto valor de $X = x$. Lo definimos como:

$$E[Y|x] = \sum y_k f_{Y|X}(y_k|x) \equiv m(x)$$

- Notar que puesto que x es una variable aleatoria entonces $m(x)$ también es una variable aleatoria.

Función de Esperanza Condicional (CEF)

- Notemos que un modelo simple para explicar o predecir sería:

$$Y = E[Y|x] + e$$

donde

$$e \equiv Y - E[Y|x]$$

- Algunas propiedades del error e:
  1. $E[e|x] = 0$
  2. $E[e] = 0$

Función de Esperanza Condicional (CEF)

Propiedades del error

- $E[e|x] = 0$

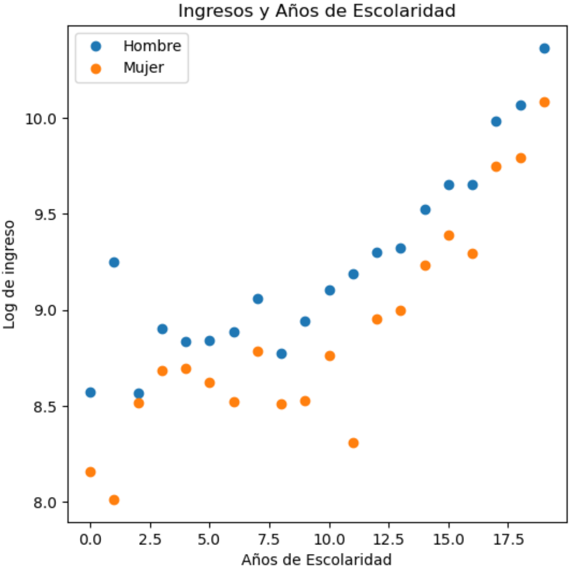
$$E[e|x] = E[y - m(x)|x] = E[y|x] - E[m(x)|x] = m(x) - m(x) = 0$$

- $E[e] = 0$

$$E[e] = E[E[e|x]] = E[0] = 0$$

Funcion de Esperanza Condicional

Aplicacion: Ingresos y Años de Escolaridad



Problema de Predicción

- La función CEF tiene una propiedad teórica interesante: provee la mejor predicción en un sentido específico.
- Supongamos que dado un vector de caracteristicas x queremos buscar una funcion $g(x)$ que nos haga la mejor predicción posible sobre y. Una forma de definir mejor predicción, es pedir que minimice el error cuadrático esperado

$$E[(y - g(x))^2]$$

Problema de Predicción

$g(x) = E[y|x] \equiv m(x)$ como la solución

- Se puede demostrar que la función que minimiza el error cuadrático medio es $g(x) = E[y|x]$.

Problema de Predicción

$g(x) = E[y|x] \equiv m(x)$ como la solución

- Una desventaja es que no siempre será fácil estimar $E[y|x]$, por ejemplo por tener pocos datos para nuestro x de interés.
- Tampoco conocemos la forma funcional de $E[y|x]$

Varianza Condicional

Nuestro objeto de interés es más que el valor esperado

- Definimos varianza condicional en general como:

$$Var(w|x) = E[(w - E[w|x])^2]$$

- Se sigue que la *varianza condicional del error del modelo CEF* es la esperanza condicional del error al cuadrado:

$$\sigma^2(x) = Var(e|x) = E[(e - E[e])^2] = E[e^2|x]$$

Y definimos tambien el *desvío estandar condicional*:

$$\sigma(x) = \sqrt{E[e^2|x]}$$

Notar que la varianza del error *no-condicional* es el valor esperado de la varianza condicional

$$\sigma^2 = E[e^2] = E[E[e^2|x]] = E(\sigma^2(x))$$

Modelo Lineal (Proyección Lineal)

- Un *caso particular* de un CEF es cuando el valor esperado cumple que puede expresarse como una función lineal:

$$m(x) = x_1\beta_1 + x_2\beta_2 + \ldots + \beta_k$$

- Para la notación es útil resumir esta forma si usamos

$$\mathbf{x} = \left\{ \begin{array}{c} x_1 \\ x_2 \\ \ldots \\ x_{k-1} \\ 1 \end{array} \right\} \boldsymbol{\beta} = \left\{ \begin{array}{c} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_{k-1} \\ \beta_k \end{array} \right\}$$

entonces

$$m(x) = \boldsymbol{x}'\boldsymbol{\beta}$$

Modelo CEF lineal o Regresion Lineal

- Por lo tanto el modelo de Regresion Lineal queda definido por:

$$y = \boldsymbol{x'}\beta$$

$$E[e|\boldsymbol{x}] = 0$$

- Usar este modelo en vez del CEF implica de hecho optar por una proyección lineal (para cualquier x).
- Nuestro modelo pierde flexibilidad, pero también tiene ventajas: nos permite realizar predicciones donde no tenemos datos de x.

Modelo CEF lineal o Regresion Lineal

- Un supuesto adicional define el Modelo de Regresion Homoscedastico:
$$E[e^2|\boldsymbol{x}] = \sigma^2$$
(La varianza es constante independiente de x)
- Notar que este supuesto se utiliza para simplificar el análisis de los modelos, pero no es algo que podemos suponer en general. Al contrario, en general esperamos que exista variabilidad condicional a x.

Encontrando el Mejor Predictor Lineal

En general, también existe un mejor modelo lineal, en el sentido que existe un modelo lineal que minimiza el error cuadrático medio.

Proof.

$$argmin_{b \epsilon \Re^k} S(\beta) = E[(y - x'\beta)^2]$$

$$= E[y^2] - 2\beta' E[xy] + \beta' E[xx']\beta$$

$$0 = \frac{\partial S(\beta)}{\partial \beta} = -2E[xy] + 2E[xx']\beta$$

$\square$

$$\Rightarrow \beta = (E[xx'])^{-1} E[xy]$$

Completamos el Modelo de Proyeccion Lineal

Remark

El modelo lineal de menor error será:

$$y = \boldsymbol{x'\beta}$$

$$E[x'e] = 0$$

$$\beta = (E[xx'])^{-1}E[xy]$$

Variables Categoricas y Dummys

Si los regresores en x toman un set finito de valores, entonces el CEF se puede escribir como un modelo lineal. En otras palabras, una regresión lineal con un número finito de valores de x (i.e. categorías) coincide con la estimación del CEF. Supongamos

$$E[y|\text{sexo}) = \begin{cases} \mu_0 & \text{si sexo=hombre} \\ \mu_1 & \text{si sexo=mujer} \end{cases}$$

Variables Categoricas y Dummys

Definimos

$$x_1 = \begin{Bmatrix} 1 & \text{si sexo=hombre} \\ 0 & \text{si sexo=mujer} \end{Bmatrix}$$

$$E[y|x] = \beta_1 x_1 + \beta_2$$

Notar que $\beta_1 = \mu_0 - \mu_1$ y $\beta_2 = \mu_1$

Variables Categoricas, Dummys y Modelos no-lineales

Incorporando más de una característica y la posibilidad de interacciones

Supongamos

$$E[y|\text{sexo}) = \begin{Bmatrix} \mu_{00} & \text{si sexo=hombre soltero} \\ \mu_{01} & \text{si sexo=hombre casado} \\ \mu_{10} & \text{si sexo=mujer casada} \\ \mu_{11} & \text{si sexo=mujer soltera} \end{Bmatrix}$$

Variables Categoricas, Dummys y Modelos no-lineales

Incorporando más de una característica y la posibilidad de interacciones

Definimos

$$x_1 = \begin{Bmatrix} 1 & \text{si casado/a} \\ 0 & \text{si soltero/a} \end{Bmatrix}, x_2 = \begin{Bmatrix} 1 & \text{si hombre} \\ 0 & \text{si mujer} \end{Bmatrix}$$

$$E[y|x_1] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4$$

Notar que $\beta_4 = \mu_{11}$, $\beta_2 = \mu_{00} - \mu_{11}$, $\beta_3 = \mu_{01} - \mu_{00} - \mu_{10} - \mu_{11}$, $\beta_1 = \mu_{10} - \mu_{11}$,

CEF y Linealidad

- El CEF es lineal, siempre y cuando las variables explicativas tomen un número finito de categorías.
- Cuando tengo una variable con I categorías puedo reducirlo a un CEF líneal, siempre y cuando traduzca las categorías en $I - 1$ variables dummies.
- Algunas variables que toman un número no muy grande de valores pueden categorizarse. En ese caso una proyección lineal y el CEF serían equivalente.

Proyeccion Lineal Vs CEF. Ejemplos

1. Modelo con Interaccion vs. sin interaccion. Caso Sexo y condicion de inmigrante

El CEF es equivalente a la especificación con interacciones:

```
. regress logingreso mujer inmigrante mujerinmigrante
```

Source	SS	df	MS
Model	369.089695	3	123.029898
Residual	7129.97081	10,107	.705448779
Total	7499.06051	10,110	.741746835

Number of obs	=	10,111
F(3, 10107)	=	174.40
Prob > F	=	0.0000
R-squared	=	0.0492
Adj R-squared	=	0.0489
Root MSE	=	.83991

logingreso	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mujer	-.295399	.0194459	-15.19	0.000	-.3335168	-.2572813
inmigrante	-.2435053	.0280165	-8.69	0.000	-.2984232	-.1885875
mujerinmigrante	-.0276682	.0381717	-0.72	0.469	-.1024924	.047156
_cons	9.50513	.0140199	677.97	0.000	9.477648	9.532612

En el caso de la mujer inmigrante (versus otras mujeres) deberia sumarse -0.02 a los -0.24 del efecto inmigrante. Un total de -0.26.

Proyeccion Lineal Vs CEF. Ejemplos

1. Modelo con Interaccion vs. sin interaccion

Al estimar sin interacción obtenemos:

```
. regress logingreso mujer inmigrante
```

Source	SS	df	MS
Model	368.719064	2	184.359532
Residual	7130.34144	10,108	.705415655
Total	7499.06051	10,110	.741746835

Number of obs	=	10,111
F(2, 10108)	=	261.35
Prob > F	=	0.0000
R-squared	=	0.0492
Adj R-squared	=	0.0490
Root MSE	=	.83989

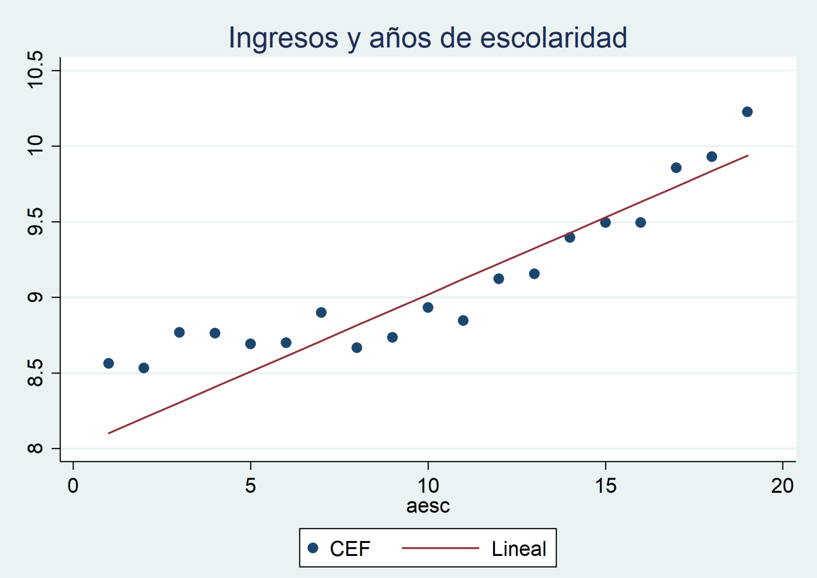
logingreso	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mujer	-.3025795	.016733	-18.08	0.000	-.3353795	-.2697795
inmigrante	-.25841	.0190282	-13.58	0.000	-.295709	-.221111
_cons	9.508862	.0130398	729.22	0.000	9.483302	9.534423

La intuición es que la proyeccion lineal promedia los casos que no surgen en la interaccion. En este caso el efecto inmigrantes (sin diferenciar sexo es -0.25, aun cuando ya incorporamos sexo como explicativa)

Proyeccion Lineal Vs CEF. Ejemplos

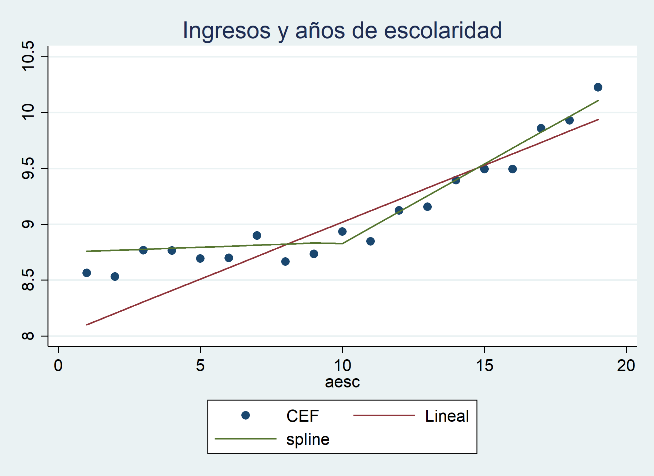
2. Lineal cuando el ajuste no es bueno

A veces la relacion lineal es buena solo en un segmento



Proyeccion Lineal Vs CEF. Ejemplos

2. Lineal cuando el ajuste no es bueno



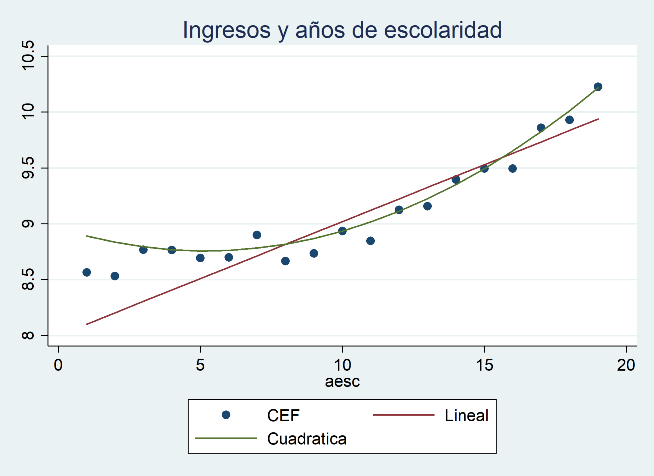
$$P(log(ingreso)|esc) = \beta_1 + \beta_2 esc + \beta_3(esc - 9) * 1(esc >= 9)$$

Proyeccion Lineal Vs CEF. Ejemplos

3. Proyeccion Cuadratica

$$P(log(ingreso)|experiencia) = \beta_1 + \beta_2 experiencia$$

$$P(log(ingreso)|experiencia) = \beta_1 + \beta_2 experiencia + \beta_3 experiencia^2$$



Práctica

Usando datos de la EAH CABA:

1. Analizar (gráficamente) la distribución del ingreso y
2. Aproximar el CEF $E[y|escolaridad]$
3. Aproximamos el CEF $E[y|escolaridad, sexo]$
4. Graficar $E[y|escolaridad]$ v.s. su mejor Proyeccion Lineal
5. Estimar y graficar una proyeccion cuadrática para $E[y|escolaridad]$
6. Estimar y graficar un modelo tipo spline.